JOURNAL OF VIROLOGY, May 1999, p. 4427–4432
0022-538X/99/$04.00+0
Vol. 73, No. 5

NOTES

Molecular Cloning and Phylogenetic Analysis of Human Immunodeficiency Virus Type 1 Subtype C: a Set of 23 Full-Length Clones from Botswana

VLADIMIR A. NOVITSKY,[1] MONTY A. MONTANO,[1] MARY F. McLANE,[1] BORIS RENJIFO,[1] FREDRIK VANNBERG,[1] BRIAN T. FOLEY,[2] THUMBI P. NDUNG'U,[1] MAFIZUR RAHMAN,[3] MOEKETSI J. MAKHEMA,[4] RICHARD MARLINK,[1] AND MAX ESSEX[1]*

Harvard AIDS Institute, Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts 02115,[1] and Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, New Mexico 87545,[2] and AIDS/STD Unit[3] and Princess Marina Hospital,[4] Gaborone, Botswana

Received 16 October 1998/Accepted 4 February 1999

To better understand the virological aspect of the expanding AIDS epidemic in southern Africa, a set of 23 near-full-length clones of human immunodeficiency virus type 1 (HIV-1) representing eight AIDS patients from Botswana were sequenced and analyzed phylogenetically. All study viruses from Botswana belonged to HIV-1 subtype C. The interpatient diversity of the clones from Botswana was higher than among full-length isolates of subtype B or among a set of full-length HIV-1 genomes of subtype C from India (mean value of 9.1% versus 6.5 and 4.3%, respectively; $P < 0.0001$ for both comparisons). Similar results were observed in all genes across the entire viral genome. We suggest that the high level of HIV-1 diversity might be a typical feature of the subtype C epidemic in southern Africa. The reason or reasons for this diversity are unclear, but may include an altered replication efficiency of HIV-1 subtype C and/or the multiple introduction of different subtype C viruses.

The majority of new human immunodeficiency virus (HIV) infections in the global AIDS epidemic are appearing in sub-Saharan Africa and Southeast Asia. Compared with the situation a decade ago, the main AIDS epidemics have shifted from central and eastern Africa to the southern regions. The most severe HIV epidemics have recently afflicted such southern African countries as Zimbabwe, Zambia, Namibia, South Africa, and Botswana (43). HIV-1 subtype C has been estimated to account for 48% of HIV-1 infections worldwide and 51.5% of HIV-1 infections in Africa (4, 7, 14–16, 21, 31), where the main mode of transmission is heterosexual (43, 44, 47).

A rapid expansion of the HIV-1 epidemic in Botswana has occurred since the early to mid 1990s. According to the UNAIDS and World Health Organization (WHO) Global HIV/AIDS & STD Surveillance data, HIV prevalence among antenatal clinic attendees tested in the major urban areas of Botswana (Gaborone, Francistown, and Selebi-Phikwe) increased from 6% in 1990 to 39% in 1997 (range of 34 to 43%) (42). Among women 20 to 29 years of age, 43 to 44% tested HIV positive. Outside of the major urban areas, median HIV prevalence increased from no evidence of infection in 1985 to 1987 to 34% in 1997. In 1997, HIV prevalence in Botswana ranged from 28 to 38%. As such, locally circulating HIV-1 needs to be characterized thoroughly, and vital information about the nature of the epidemic should be extended (2, 4, 7, 21, 31, 37, 45–47). Moreover, Botswana's central geographic position makes a comprehensive HIV-1 molecular epidemiological study that much more urgent, because it may serve as an example of the burgeoning epidemic in southern Africa.

In this study, we report the molecular cloning and phylogenetic analysis of 23 near-full-length clones from Botswana. All of them were identified as belonging to HIV-1 subtype C and demonstrated high levels of intersample diversity across the entire viral genome. By providing new genetic information regarding locally circulating viruses, this study may contribute to AIDS vaccine design for the southern Africa region countries and, in particular, for Botswana.

Specimens for this study were selected from HIV-seropositive patients in Gaborone, Botswana. All HIV-1 infections in this study were likely to be heterosexually acquired. The times of infection were not known. The HIV-1-seropositive status of patients was confirmed by enzyme-linked immunosorbent assay and Western blot analysis. Clinical classification was performed by using the 1987 Centers for Disease Control and Prevention (CDC) revised classification (9) (data not shown).

Genomic DNA was obtained directly from the patients' peripheral blood mononuclear cells (PBMCs)—buffy coats—without passage through cell culture or donor PBMCs. All clones in this study were amplified in heminested PCR with three primers from the LA set (18) or their modifications. The Expand Long Template PCR system (Boehringer Mannheim, Indianapolis, Ind.) was used according to the manufacturer's instructions. Gel purification of the first-round PCR product was essential for direct amplification of 9.0-kb fragments from uncultured PBMCs. Estimation of the expanded PCR sensitivity (based on 8E5/LAV) revealed a successful amplification of the 9.0-kb fragment in the first round when at least 8×10^2 to

* Corresponding author. Mailing address: Department of Immunology and Infectious Diseases, Harvard School of Public Health, FXB-402, 651 Huntington Ave., Boston, MA 02115. Phone: (617) 432-0975. Fax: (617) 739-8348. E-mail: messex@sph.harvard.edu.

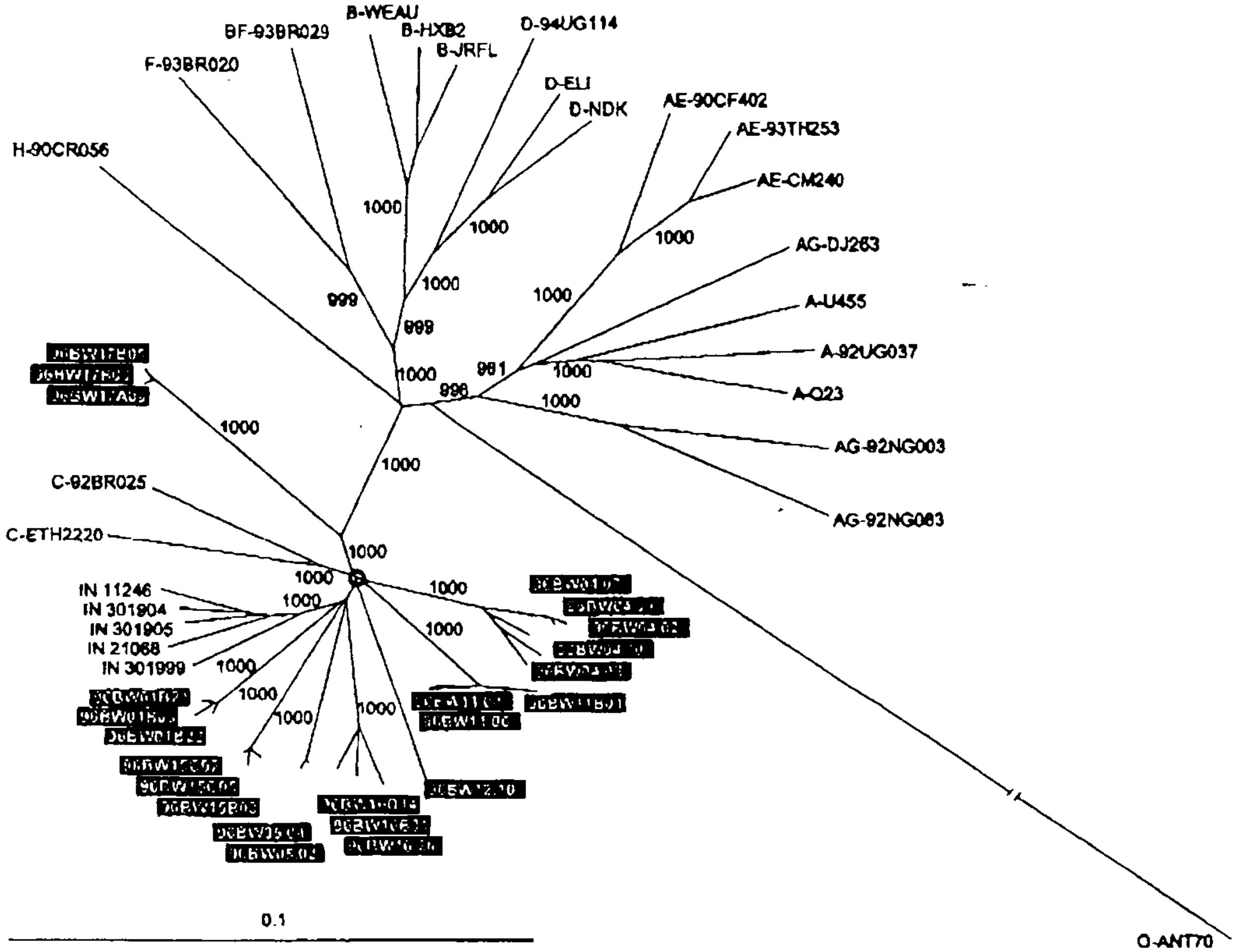

FIG. 1. Phylogenetic relationship of the newly characterized full-length clones from Botswana (boxed in black) to other representative full-length HIV-1 sequences of subtypes A, B, C, D, F, and H and recombinant subtypes AE, AG, and BF. Full-length subtype C sequences from India were also included in the analysis. A neighbor-joining tree was constructed on the basis of the hidden Markov model nucleotide alignment of full-length HIV-1 genomes. Subtype O ANT70 sequence was used as an outgroup. Values along the branches indicate the bootstrap values that support branching (out of a 1,000 resampling).

4 × 10^3 proviral copies were present in the reaction (data not shown). These results were consistent with those from other studies (10, 39). The TA pCR2.1 TOPO system (Invitrogen, Carlsbad, Calif.) and JM109 competent cells (Promega Corporation, Madison, Wis.) were used for cloning. Positive colonies were screened by PCR. To obtain sufficient plasmids for sequence analysis, we amplified the constructs under the previously described conditions with some modifications (41). Purified plasmid DNA served as a template for sequencing. Both-strand sequencing was combined with a strategy involving overlapping sequences. Dye terminator sequencing on an automated DNA Sequenator (model 373A; Applied Biosystems, Inc., Foster City, Calif.) was used.

A multiple alignment procedure for the full-length HIV genome was performed by using the hidden Markov model. Constructed through the HIV-1 HMMER computer program of the Los Alamos National Laboratory, the model has been previously shown to provide the best description of the true nucleotide substitution pattern of HIV-1 *gag* and *env* genes (26). The HIV-1 HMMER model (11, 12) constructed at Los Alamos National Laboratory for the full-length HIV-1 genomes (24) was employed. Sixty full-length reference sequences were included in the alignment from the GenBank data bank (5). The 3′ end of the alignment, which included the *nef* coding region and 3′ long terminal repeat (LTR), was adjusted manually. The pairwise evolutionary distances from nucleotide sequences were computed by the DNADIST program under Kimura's two-parameter model (17). All alignments were globally gap stripped for the generation of the trees. The transition/transversion ratio parameters were set at 3.0 for the *gag* gene, 1.5 for the *env* gene, 1.42 for the V1-V2 and V3 fragments, and 2.0 for the other viral loci (25). A tree was drawn by the Njplot (33) and TreeView (32) programs. To analyze patterns of variability along the HIV-1 genomes, the program SWAN, which utilizes a "sliding window" approach was used (34). Positions with gaps either were or were not excluded from the analysis. The variability distribution was estimated as an entropy function of the nucleotide variation observed at a particular position. The Recombinant Identification Program (RIP) (40) and HIV-1 Subtyping Basic BLAST (3) were used in searching for recombination among the clones studied.

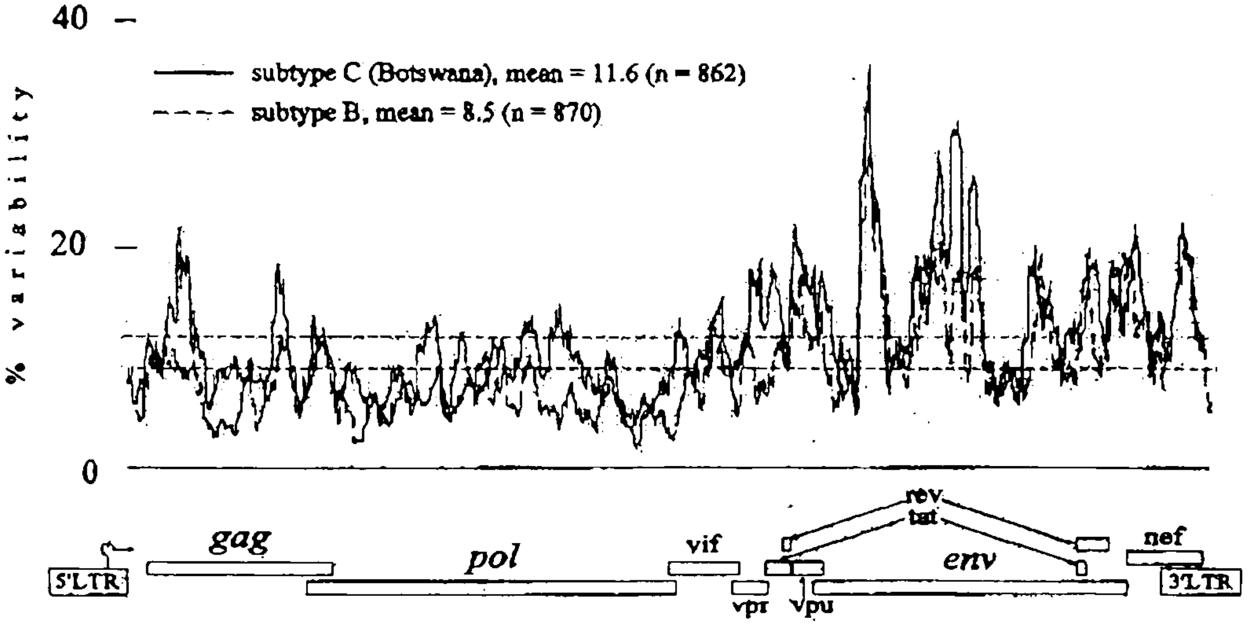

FIG. 2. Variability plots comparing subtype B and C (Botswana) sequences across the entire HIV-1 genome. The variability distribution was estimated as an entropy function of the nucleotide variation in the SWAN program based on the hidden Markov model alignment of the complete HIV-1 genome. The subtype B sequences used in the analysis were DH123, 89.6, RF, WEAU, OYI, HXB2, JRFL, and YU2. *n*, number of sliding window sites across the HIV-1 genome with gap stripping.

Sequence analysis of the Botswana HIV-1 revealed that 10 of 23 clones had an intact genomic organization with open reading frames. The other clones had point mutations and/or insertions and deletions resulting in frameshifts, disabled start codons, or premature stop codons. No major deletions or rearrangements were observed. Determined length polymorphism among studied sequences was limited to the *vpu* (15- to 18-nucleotide [nt] insertion at the 5' end), *env* (from 3- to 9-nt deletions to 48-nt insertions, GIGRGQ motif in the BW17 V3 loop), 2nd exon of *rev* (a 13-amino-acid truncation at the 3' end), *nef* (6- to 15-nt deletions in some clones), and regulatory regions of the LTR (three or four NF-κB sites with GGGAC TTTCT as a potential fourth NF-κB in two clones of BW05).

An evolutionary tree in Fig. 1 shows the phylogenetic relationship of the full-length Botswana clones to other representative full-length HIV-1 sequences. All Botswana sequences, a set from India (27), and two subtype C reference sequences (C-ETH2220 [38] and C-92BR025 [19]) clustered together, forming a compelling subtype C outcropping on the phylogenetic tree. This cluster was separated from other HIV-1 sequences by the extremely high bootstrap value of 1,000 (out of 1,000 resampling). Phylogenetic relationships within the subtype C bush were also noteworthy. Assuming that the circled node at the center of the bush could represent the potential ancestral subtype C node, we observed the following. (i) The star-like phylogeny of the subtype C bush together with its branching order may demonstrate the relatively high diversity of the Botswana samples. (ii) Four Botswana sample clones (BW01, BW05, BW15, and BW16), together with all five sequences from India, formed a potential subcluster, although the bootstrap value was not high. (iii) All Indian samples were separated by bootstrap values of 1,000, possibly reflecting a "founder effect" among these sequences. (iv) Three Botswana sample clones (BW04, BW11, and BW12) may represent individualized groups of sequences inherited from a common subtype C ancestor. (v) Two reference sequences ETH2220 and 92BR025 deviated together with a high bootstrap value (1,000), possibly reflecting another subtype C subcluster differing significantly from Botswana or Indian samples. (vi) One of the Botswana samples (BW17) strayed rather far off the main subtype C bush and may be the least representative of Botswana HIV-1 samples. (vii) The topology of the Botswana clones confirms that clones of the same samples are closely related to each other based on full-length genome sequences. A multilocus analysis was congruent with full-genome phylogenetic analysis and confirmed clustering of newly derived Botswana clones within subtype C across the entire HIV-1 genome (data not shown).

To characterize the level of variability among Botswana clones across the entire HIV-1 genome, we performed SWAN program analysis as an entropy function of the nucleotide variation. The Botswana set had greater variability than subtype B samples (Fig. 2) (mean values of 11.6 and 8.5%, respectively, for gap-stripped analysis). The profiles of viral variability across the HIV-1 genome were similar among subtype B and C viruses. Comparison of gap-stripped and gap-nonstripped plots revealed that the differences in mean values between the two methods of computing and the shape of variability plot profiles were not significant (data not shown). Gap stripping slightly decreased the mean value and the number of sliding window sites across the genome. It also hid the extreme regions with the highest level of variability. Both variability when measured as an entropy function in this study and when described before diversity as a pairwise comparison of the sequence (19) exhibited similar profiles of variable and conservative genomic regions. Variability plots (especially non-gap stripped) revealed higher peaks in variable regions than diversity plots.

Table 1 depicts the high degree of intersample diversity across the entire HIV-1 genome among Botswana clones compared with subtype B and C sequences from India (27). Be-

TABLE 1. Intersample HIV-1 diversity in this study[a]

Gene or region	Mean % (range) HIV-1 diversity[b]			
	Subtype C, Botswana	Subtype B		Subtype C, India
		AIDS patients (8 sequences)	23 isolates	
gag	7.9 (5.9–9.2)	5.6* (3.6–9.6)	4.9* (1.0–8.6)	3.4* (2.4–4.6)
pol	5.9 (4.3–7.8)	4.1* (2.9–5.4)	3.9* (0.8–6.2)	2.6* (1.5–3.7)
vif	7.5 (4.3–12.6)	6.3† (3.7–9.3)	6.4‡ (0–9.5)	3.2* (1.8–4.3)
vpr	9.8 (5.3–14.9)	6.2* (4.3–8.4)	5.9* (1.7–10.8)	4.8* (2.1–8.1)
tat	9.9 (6.9–14.9)	7.5* (4.6–10.7)	7.1* (3.4–10.7)	3.7* (2.7–5.1)
rev	10.4 (7.7–16.6)	9.4§ (4.2–14.2)	8.2* (3.3–14.2)	4.7* (2.8–5.7)
vpu	13.9 (10.1–18.8)	10.8* (6.1–14.3)	9.5* (4.2–15.8)	5.1* (2.1–8.1)
env	12.4 (10.4–14.5)	9.5* (6.2–11.9)	9.5* (5.9–12.7)	6.9* (5.1–9.3)
V1-V2	25.7 (15.9–36.7)	18.0* (8.6–26.1)	19.0* (8.6–30.8)	25.8 (14.2–34.8)
V3	14.4 (11.5–18.9)	11.6* (6.7–17.3)	12.1* (6.0–18.2)	6.4* (5.1–7.9)
nef	11.3 (8.0–15.1)	9.8‖ (6.3–15.7)	9.2* (3.5–16.3)	5.2* (4.4–6.4)
3′ LTR	9.7 (6.6–13.4)	8.3# (5.1–11.6)	7.3* (1.0–11.6)	4.0* (3.2–5.4)
Full-length genome	9.1 (7.7–10.7)	6.6* (4.5–8.0)	6.5* (3.5–9.6)	4.3* (3.2–5.7)

[a] Fifty-six full-length HIV-1 sequences were used in the analysis, based on the hidden Markov model alignment of the entire HIV-1 genome. The following 26 sequences of subtype B were used: AUMBCC54, C18MBC, DH123, 89.6, RF, WEAU, HAN, HIVMN, BCSG3, OYI, CAM1 NY5, pNL43, LAI, HXB2, JRFL, JRCSF, AUMBC925, AUMBC200, YU2, YU10, ACH320A, ACH320B, SF2, HIV1AD8, D31, MANC, and WR27. Sequences AUMBCC54, C18MBC, and NY5 were excluded from the *nef* and 3′ LTR analysis because of deletions or the absence of sequences for these regions. Pairwise distances in four groups of sequences (pNL43, LAV, and HXB2; JRFL and JRCSF; YU2 and YU10; and ACH320A and ACH320B) were excluded from the analysis. The eight subtype B sequences from AIDS patients were JRFL, YU2, 89.6, RF, HAN, MN, BCSG3, and WR27. Thirteen sequences of subtype C were included in the analysis: 8 clones from Botswana (1 from each patient) and 5 sequences from India (301999, 21068, 301905, 301904, and 11746). All distances were calculated by DNADIST program from the PHYLIP v. 3572 package based on hidden Markov model alignment. The transition/transversion ratios were set to 3.0 for *gag*, 1.5 for *env*, 1.42 for V3 and V1-V2, and 2.0 for all other HIV-1 genes.

[b] Statistical significance versus Botswana HIV-1 clones is shown as follows: *, $P < 0.0001$; #, $P = 0.005$; ‖, $P = 0.013$. ‡, $P = 0.028$; †, $P = 0.033$; and §, $P = 0.13$.

cause AIDS patients might be expected to have higher variability, we made the same comparison, limiting the subtype B reference to eight sequences selected from confirmed AIDS patients (column 2). The intersample diversity among Botswana clones was significantly higher than that among subtype B references or Indian samples on the level of the full-length HIV-1 genome. Across the viral genome, the mean diversity among Botswana samples was congruent with the full-length genome analysis. The intersample diversity analysis statistically confirmed the phylogenetic study observations (Fig. 1 and 2) that the newly characterized Botswana clones were highly diversified.

The results of intrasample diversity analysis were limited by the methods used (PCR amplification and cloning) and by the available number of multiple subtype B full-length clones. Seven Botswana samples (except BW12) and three subtype B sets (JRFL, YU, and ACH320) were compared across the HIV-1 genome. The range of full-length diversity among Botswana samples was 0.3% (BW17 clones) to 2.9% (BW04 clones), with an average mean value of 1.4%. Intrasample diversity showed no significant difference between subtype B and C sequences (Fig. 3). However, two concentrations of diversity (low and high) were revealed among both subtype B and C sets (Fig. 3). These concentrations of low and high diversity were distributed across the entire genome and were found to be more consistent in the structural genes (*gag*, *pol*, and *env*).

All Botswana sequences were checked for potential recombination sites by the HIV-1 Subtyping Basic BLAST (3) and by RIP (40), the results consistently showing no evidence of recombination.

Clustering with HIV-1 subtype C and the high intersample diversity were the most exceptional attributes of the 23-clone set from Botswana. A star-like shape of the subtype C cluster in the phylogenetic tree was accompanied by extremely high bootstrap values across tree branches. The topology of the phylogenetic tree suggested that a common ancestor for the Botswana sequences might have existed before the common ancestor for the Indian sequences analyzed or before the strains C-92BR025 (Brazil) and C-ETH2220 (Ethiopia) diverged.

Intersample diversity within subtype C has been previously found to vary from 5 to 11.5% (1, 7, 38). Higher levels of diversity were found among Botswana clones in this study, in spite of the fact that samples were taken from one place and at one time point. Both full-genome sequences and multiple subgenomic loci demonstrated the same patterns, with a higher mean value of variability among the Botswana samples.

The increased genetic diversity of subtype C viruses in Botswana might have different underlying causes, including a variety of host and viral factors. Among the latter factors, a combination of the genetic flexibility of subtype C virus and its multiple introductions might be the most important. A number of recent findings argue that one possible cause of the high viral diversity in the Botswana epidemic could be higher flexibility of subtype C virus and its altered ability to diversify. These arguments include, but are not limited to the following. (i) Subtype C is predominant in most recent HIV-1 epidemics worldwide (1, 7, 27, 35, 36, 38, 45, 47). (ii) The highest prevalence of HIV-1 infection in various epidemics is caused mainly by subtype C virus. (iii) Subtype C virus may have a faster disease progression (20), and patients infected with HIV-1 subtype C developed AIDS earlier than patients with subtype A virus (23). (iv) Three or four NF-κB sites (instead of two) might lead to more efficient viral transcription (13, 29, 30). (v) The TNF-α response to subtype C virus is significantly higher than to HIV-1 subtype B (28, 29), suggesting the possibility of increased viral transcription and replication in correlation with NF-κB copy number (28, 29). (vi) The viral load of subtype C infections may be higher in different compartments that might cause an increased level of viral transmission (22). On the other hand, a scenario that suggests independent diversification of the virus in other regions and delayed entry of the epidemic in Botswana, followed by multiple introductions of the subtype C virus from adjacent countries cannot be excluded (42–44).

Botswana is geographically located at the center of the AIDS epidemic in Southern Africa. UNAIDS and World Health Organization surveillance data suggest that the widespread rise of the HIV-1 epidemic in Botswana started in the early to mid-1990s and reached one of the highest prevalence rates in Africa (42–44). For more recent HIV-1 epidemics, such as those described in Thailand and India, one might expect to find a highly homogeneous pool of local viruses that formed a monophyletic phylogenetic subcluster with relatively short and aggregated branches. However the findings in this study contradict the established trend.

Extremely high interpatient diversity across the genome was supported by long branch lengths in the phylogenetic trees throughout the Botswana viruses within the genetic subtype C. No multiple subtypes or recombination were found in this study. Because it currently has the highest incidence rates of